Our Ref.: 51026.P010

APPLICATION FOR UNITED STATES LETTERS PATENT

FOR

Determining A Rating For A Collection Of Documents

Inventor(s):
John T. Larason
Alan J. Packer

Prepared by:
Columbia IP Law Group, PC

"Express Mail" label number EL910784231US

Determining A Rating For A Collection of Documents

This application claims priority to provisional application numbers 60/289,587, 60/289,400 and 60/289,418, all filed on May 7, 2001, entitled "Method of Assigning Ratings to Collections of Related Objects", "Method and Apparatus for Automatically Determining Salient Features for Object Classification" and "Vvery-Large-Scale Automatic Categorizer For Web Content" respectively having at least partial common inventorship as the present application.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to the field of data processing. More specifically, the present invention relates to automated methods and systems for determining a rating for a rating scale for a collection of documents.

2. Background Information

The World Wide Web (WWW) is an expanding collection of textual and non-textual material which is available for access to any Internet user, from any location at any time. Some users find particular contents to be objectionable. For example, parents often wish to shield their children from exposure to sexually explicit material, hate speech, and drug information. Similarly, companies may wish to prevent access by employees to web sites that provide or support gambling.

Notwithstanding the civil liberty implications associated with these concerns, a number of groups and companies have brought forward systems and techniques for assisting Internet users in block accessing to undesired content. For example,

various blocking software products are available from software vendors, such as SafeSurf of Newbury Park, CA, and NetNanny of Bellevue, WA. Typically, these products employ site lists to effectuate blocking of access to undesired contents. These site lists include the identifications of the web sites containing undesired contents. Access to any of the web pages hosted by the identified web sites is blocked. Another example of such a system is described by Neilsen et al., "Selective downloading of file types contained in hypertext documents transmitted in a computer controlled network", US Patent 6,098,102, which utilizes the file extensions of URLs to determine whether the particular files will or will not be downloaded to the user. Still another method for controlling access to web sites is typified by the work of the Internet Content Rating Association, which uses the technology of the Platform for Internet Content Selection (PICS) specification to allow voluntary, or in the future potentially mandatory, rating of page content by the content author. Filtering can then be done by utilizing these rating "tags", and may be augmented by a complete block on other un-rated pages.

These prior art approaches suffer from at least the following disadvantages:

a) The WWW is constantly growing. The number of web sites and their contents are constantly changing. As a result, the prior art approaches are unable to keep pace with the changes.

b) Further, many web sites generate user-specific pages at every access. As a result, the prior art URL based approaches are unable to facilitate blocking of these dynamically generated pages if they contain undesired contents.

c) Additionally, content providers are often not the best, or even the appropriate, agent for rating their own contents. Duplicitous providers may deliberately mis-rate the appropriateness of their contents.

Some filtering systems rely on key word lists or text analysis, to judge the content of individual pages. While these systems may work satisfactorily on text files, they are ineffective for non-text materials, such as images, sound files, or movies.

Thus, an improved approach for blocking undesired contents is desired.

SUMMARY OF THE INVENTION

On one or more data processing systems, a collection rating is determined for a rating scale for contents of a document collection. A link rating is determined for the rating scale for contents linked to or linked by contents of the document collection. The collection rating for the rating scale for contents of the document collection is then modified, based on the determined link rating for the rating scale for contents linked to or linked by contents of the document collection.

In one embodiment, a collection rating for a rating scale for a document collection is determined based on document ratings of a subset of the documents of the document collection, and their sizes.

In one embodiment, the link rating for the rating scale for the document collection is determined based on the collection ratings of the document collections having contents linked to or linked by contents of the document collection.

In one embodiment, the document collection is a web site, the documents of the document collection are web pages of the web site, and the subset of documents employed to determine the web site rating is the textual documents.

Note: The term "document" as used herein in this application, including the specification and the claims, includes textual as well as non-textual documents, unless one or more types of "documents" are expressly excluded or implicitly excluded in view of the context of the usage.

BRIEF DESCRIPTION OF DRAWINGS

The present invention will be described by way of exemplary embodiments, but not limitations, illustrated in the accompanying drawings in which like references denote similar elements, and in which:

Figure 1 illustrates an overview of the present invention in accordance with one embodiment;

Figure 2 illustrates a method view of the present invention, in accordance with one embodiment;

Figure 3 illustrates the operational flow for determining a collection rating, in accordance with one embodiment;

Figure 4 illustrates the operational flow for determining a link rating, in accordance with one embodiment; and

Figure 5 illustrates a computer system suitable for use to practice the present invention, in accordance with one embodiment.

Glossary

URL – Uniform Resource Locator

DETAILED DESCRIPTION OF THE INVENTION

As summarized earlier, the present invention includes improved methods and related apparatuses for determining a rating for a rating scale for a document collection. In the description to follow, various aspects of the present invention will be described. However, the present invention may be practiced with only some or all aspects of the present invention. For purposes of explanation, specific numbers, materials and configurations are set forth in order to provide a thorough understanding of the present invention. However, the present invention may be practiced without some of the specific details. In other instances, well known features are omitted or simplified in order not to obscure the present invention.

Parts of the description will be presented in terms of operations performed by a processor based device, using terms such as data, analyzing, assigning, selecting, determining, and the like, consistent with the manner commonly employed by those skilled in the art to convey the substance of their work to others skilled in the art. As well understood by those skilled in the art, the quantities take the form of electrical, magnetic, or optical signals capable of being stored, transferred, combined, and otherwise manipulated through mechanical and electrical components of the processor based device. The term "processor" includes microprocessors, micro-controllers, digital signal processors, and the like, that are standalone, adjunct or embedded.

Various operations will be described as multiple discrete steps in turn, in a manner that is most helpful in understanding the present invention. However, the order of description should not be construed as to imply that these operations are

necessarily order dependent. In particular, these operations need not be performed in the order of presentation. Further, the description repeatedly uses the phrase "in one embodiment", which ordinarily does not refer to the same embodiment, although it may.

Overview

Referring now to **Figure 1**, wherein a block diagram illustrating an overview of the present invention, in accordance with one embodiment, is shown. As illustrated, collection rater **110** of the present invention, is equipped to deduce a collection rating **112** for a rating scale for a document collection, such as collection **102**. An example of a rating scale is a scale that quantitatively rates the contents of a subject collection on its "offensiveness", e.g. ranging from 0 to 3, with 0 meaning "not offensive", 1 meaning "mildly offensive", 2 meaning "moderately offensive" and 3 meaning "very offensive". As will be described in more detail below, collection rater **110** advantageously generates collection rating **112** for a collection taking in account not only the contents of the collection, but also contents of other collections linked to or linked by contents of the subject collection, such as collection **104** and collection **106** respectively. As those skilled in the art would appreciate, the inclusion of the contents linked to or linked by contents of the subject collection tends to strengthen the accuracy of the rating generated for the subject collection.

In one embodiment, collections **102**, **104** and **106** are web sites, and documents **103**, **105** and **107** are web pages of the web sites, including textual as well as non-textual, such as multi-media, web pages. In alternate embodiments, documents **103**, **105** and **107** may be other content objects, with collections **102**, **104** and **106** being other organizational entities of the content objects.

Method

Referring now to **Figure 2**, wherein a block diagram illustrating a method view of the present invention, in accordance with one embodiment, is shown. As illustrated, for the embodiment, collection rater **110** generates a collection rating for rating scale for a subject collection, by first determining an initial collection rating for the contents of the subject collection, block **202**. Upon so determining, collection rater **110** determines a link rating for the contents of the linked collections, i.e. collections with contents linked to or linked by contents of the subject collection, block **204**. Thereafter, for the illustrated embodiment, collection rater **110** modifies the initially determined collection rating, using the determined link rating, thereby taking into consideration the "linked" contents, block **206**.

In one embodiment, in block **206**, collection rater **110** modifies the initially determined collection rating by replacing the initially determined collection rating with the determined link rating. In another embodiment, in block **206**, collection rater **110** modifies the initially determined collection rating by adding the determined link rating to the initially determined collection rating. In yet another embodiment, in block **206**, collection rater **110** modifies the initially determined collection rating by subtracting the determined link rating from the initially determined collection rating. In yet other embodiments, in block **206**, collection rater **110** may modify the initially determined collection rating by combining the determined link rating with the initially determined collection rating in other alternate manners.

The manner in which the determined link rating is to be combined with the initially determined collection rating to modify the initially determined collection rating to take into account the linked contents is application dependent. Preferably, the manner of combination is user configurable. Such user configuration may be facilitated through any one of a number of user configuration techniques known in

the art, which are all within the abilities of those ordinarily skilled in the art. Accordingly, no further description of these user configuration techniques is necessary.

Collection Rating

Referring now to **Figure 3**, wherein a block diagram illustrating a manner collection rater **110** generates a collection rating for a rating scale for a subject collection, in accordance with one embodiment, is shown. As illustrated, for the embodiment, collection rater **110** generates the collection rating for a rating scale for a subject collection by first determining the individual document ratings for a subset of the documents of the subject collection, block **302**. In one embodiment, the subject collection comprises textual as well as non-textual, such as multi-media, documents. For the embodiment, the subset of the documents is the textual documents. The determination of the individual document ratings for the textual documents may be made in accordance with any one of a number of document rating techniques, e.g. by the salient features or keywords of each of the document. Examples of these document rating techniques include but are not limited to those described in U.S. Provisional Applications numbers 60/289,400 and 60/289,418, entitled "METHOD AND APPARATUS FOR AUTOMATICALLY DETERMINING SALIENT FEATURES FOR OBJECT CLASSIFICATION" and "VERY-LARGE-SCALE AUTOMATIC CATEGORIZER FOR WEB CONTENT" respectively, both filed on May 7, 2001. Both applications are hereby fully incorporated by reference.

In accordance with the present invention, in addition to determining the individual document ratings of the subset of the documents, collection rater **110** further determines the sizes of the documents, block **304**. Then, collection rater **110**

determines the collection rating by combining the determined individual document ratings in a size and rating normalized manner, block **306**.

More specifically, in one embodiment, collection rater **110** combines the determined individual document ratings in a size and rating normalized manner, by grouping the documents in accordance with their determined sizes and determined ratings, and applying weights to the determined document ratings in accordance with their size group and rating group membership. In one embodiment, the weights are applied in accordance with the group sizes and determined ratings as set forth by the tables below:

Document size range in (bytes)	Weight
<500	1
500 - 999	4
1000 - 4999	7
5000 - 9999	10
>9999	13

Determined document rating for said rating scale	Weight
0	-0.5
1	0.5
2	3
3	6

The weights are applied in accordance with the formula set forth below:

$$CR = \frac{\sum_{i,j} r_i w_j \log(N_{ij} + 1)}{\sum_{i,j} w_j \log(N_{ij} + 1)}$$

where *CR* is the collection rating for the rating scale;

r_i is the weight applied for document rating group *i*;

w_i is the weight applied for document size group *j*;

N_{ij} is the number of pages in the collection with document rating *i* and having group sizes *j* for the rating scale.

In alternate embodiments, for different rating scales, different rating and/or group size based weighting schemes, as well as other weighting schemes may be employed instead.

Link Rating

Referring now to **Figure 4**, wherein a block diagram illustrating a manner collection rater **110** generates a link rating for a rating scale for a subject collection, in accordance with one embodiment, is shown. As illustrated, for the embodiment, collection rater **110** generates the link rating for a rating scale for a subject collection by first generating the collection ratings for the collections having contents either linked to or linked by contents of the subject collection, block **402**. The collection rating for the rating scale for each of the collection with contents either linked to or linked by contents of the subject collection, may be generated in the same manner the collection rating for the rating scale for the subject collection is generated, e.g. as earlier described, or in a different manner.

Upon so determining, for the illustrated embodiment, collection rater **110** sums the determined collection ratings for the rating scale for the other collections, block **404**, then generates the link rating based on the resulting sum, block **406**. In

one embodiment, collection rater **110** generates the link rating based on the resulting sum in accordance with the discrete "step" function set forth below:

The resulting sum (RS)	link rating
RS less than -2	- 1.0
RS greater than or equal to –2, but less than - 1	- 0.5
RS greater than or equal to –1, but less than or equal to – 0.5	0
RS greater than –0.5, but less than or equal to 1.5	0.5
RS greater than 1.5, but less than or equal to 3	1.0
RS greater than 3, but less than or equal to 4	1.5
RS greater than 4	2.0

In alternate embodiments, the link rating may be generated from the determined collection ratings of the "linked" collections employing different functions.

Accordingly, under the present invention, "linked" contents are taken into consideration to potentially strengthen the accuracy of the rating generated for a rating scale for a subject collection. As those skilled in the art would appreciate, the present invention may be practiced for one or more rating scales on one or more subject collections, each having zero or more "linked" collections. Subject collections with zero "linked" collection is merely a degenerate case where no

"linked" content contribution can be extracted to potentially strengthen the accuracy of the ratings generated for the rating scales for the subject collections.

Example Computer System

Figure 5 illustrates an exemplary computer system **500** suitable for use to practice the present invention, in accordance with one embodiment. As shown, computer system **500** includes one or more processors **502** and system memory **504**. Additionally, computer system **500** includes one or more mass storage devices **506** (such as diskette, hard drive, CDROM and so forth), one or more input/output devices **508** (such as keyboard, cursor control and so forth) and communication interfaces **510** (such as network interface cards, modems and so forth). The elements are coupled to each other via system bus **512**, which represents one or more buses. In the case of multiple buses, they are bridged by one or more bus bridges (not shown). Each of these elements performs its conventional functions known in the art. In particular, system memory **504** and mass storage **506** are employed to store a working copy (**514a**) and a permanent copy (**514b**) of the programming instructions implementing the teachings of the present invention (collection categorizer). The permanent copy (**514b**) of the programming instructions may be loaded into mass storage **506** in the factory, or in the field, as described earlier, through a distribution medium (not shown) or through communication interface **510** (from a distribution server (not shown)). The constitution of these elements **502-512** are known, and accordingly will not be further described.

In alternate embodiments, the present invention may be practice on multiple systems sharing common and/or networked storage.

Modifications and Alterations

While the present invention has been described referencing the illustrated and above enumerated embodiments, the present invention is not limited to these described embodiments. Numerous modification and alterations may be made, consistent with the scope of the present invention as set forth in the claims to follow. Of course, the above examples are merely illustrative. Based on the above descriptions, many other equivalent variations will be appreciated by those skilled in the art.

Conclusion and Epilogue

Thus, a method and apparatus for generating a collection rating for a document collection comprising textual and non-textual documents, has been described. Since as illustrated earlier, the present invention may be practiced with modification and alteration within the spirit and scope of the appended claims, the description is to be regarded as illustrative, instead of being restrictive on the present invention.
